UCRL-JC-128485
PREPRINT

The Use of An Operational Model Evaluation System for Model Intercomparison

K.T. Foster
G. Sugiyama
J.S. Nasstrom
J.M. Leone, Jr.
S.T. Chan
B.M. Bowen

This paper was prepared for submittal to the
5th International Conference on Harmonisation within
Atmospheric Dispersion Modelling for Regulatory Purposes
Rhodes, Greece
18–21 May 1998

March 1998

DISCLAIMER

This document was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor the University of California nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or the University of California, and shall not be used for advertising or product endorsement purposes.

The Use of an Operational Model Evaluation System for Model Intercomparison

K. T. Foster, G. Sugiyama, J. S. Nasstrom, J. M. Leone, Jr., S. T. Chan, Brent M. Bowen

Atmospheric Sciences Division, Lawrence Livermore National Laboratory, USA

Abstract

The Atmospheric Release Advisory Capability (ARAC) is a centralized emergency response system used to assess the impact from atmospheric releases of hazardous materials. As part of an on-going development program, new three-dimensional diagnostic windfield and Lagrangian particle dispersion models will soon replace ARAC's current operational windfield and dispersion codes. A prototype model performance evaluation system has been implemented to facilitate the study of the capabilities and performance of early development versions of these new models relative to ARAC's current operational codes. This system provides tools for both objective statistical analysis using common performance measures and for more subjective visualization of the temporal and spatial relationships of model results relative to field measurements. Supporting this system is a database of processed field experiment data (source terms and meteorological and tracer measurements) from over 100 individual tracer releases.

1. Introduction

The Atmospheric Release Advisory Capability (ARAC), a major program within the Earth and Environmental Sciences Directorate of Lawrence Livermore National Laboratory, provides a centralized emergency response service for assessing the impact from atmospheric releases of hazardous materials (Sullivan, et al. 1993). Real-time guidance is provided to U.S. Department of Energy and Department of Defense emergency response managers in the form of computer-generated isopleths of contaminant air concentration and surface deposition. As part of an on-going development program, the first phase of a major upgrade of ARAC's operational emergency response system is nearing completion. A key portion of this work is the development of new three-dimensional diagnostic windfield and Lagrangian particle dispersion models which will soon replace ARAC's current operational windfield and dispersion codes.

Prior to the operational use of the new models, a complete study of their capabilities and performance (accuracy, robustness, efficiency, etc.) relative to ARAC's current models must be completed. As part of this effort, ARAC has developed a prototype model evaluation system for both the objective statistical analysis and the subjective visualization of the inter-relationships between the results of a model's calculations and the appropriate field measurement data.

Numerous model evaluation studies have been conducted during ARAC's operational lifetime (Foster and Dickerson, 1990). The results of these previous studies were used to evaluate the performance of ARAC's current operational diagnostic, three-dimensional, mass-consistent windfield model, MATHEW (Sherman, 1978 and Sugiyama, et al., 1994), and Lagrangian particle, random-walk dispersion model, ADPIC (Lange, 1978 and Ermak, et al., 1995). A review of the previous studies and available data sets showed that although results were typically presented within a common framework using primarily a ratio factor analysis, a strict adherence to a common set of evaluation protocols was not maintained since the work was performed by a number of scientists over a 20-year period. In some cases the measurement data and related information used in earlier studies could not be located for re-use.

A goal of the current work is to develop improved evaluation tools for studying and understanding a model's performance as it is applied to various source term, meteorological, and topographical situations. The new tools are designed to maintain a persistent database of meteorological and concentration measurements to support on-going evaluation studies.. They provide for the extraction and formatting of the data for repeated use in batch model runs which allow the user to easily execute several model simulations with different versions of the model. The principal goals of the implementation are to track statistical changes in model performance as new transport and diffusion models or algorithms are implemented, and to provide an analysis framework which permits users to arrive at a better understanding of the reasons for these changes. A new set of

MATHEW/ADPIC evaluation studies has been completed using these tools to apply common protocols and assumptions. These new studies provide a benchmark for future comparisons. Future applications are meant to serve the needs of both quality control of the ARAC operational codes and of detailed scientific investigation.

2. System Protocols

A wide range of statistical measures and protocols are used in model evaluation studies. Often the selected measures and protocols are chosen to be consistent with the application of the particular model under study, or to meet the specific goals of the study. ARAC's primary application is emergency response. Users of ARAC's modeling are emergency response managers or others interested in assessing the actual or potential impact of hazardous atmospheric releases within very short time constraints. Over the years the focus of these real-time assessments have changed from depictions of general overall contamination patterns to estimates of much more local effects. Guidance is expected to address questions of contamination at specific points of interest (e.g., schools or hospitals) and possibly at specific times (e.g., exposure of a moving population or planning for potential evacuations under extended release conditions).

Evaluation of such applications requires analyses on very localized scales. Fine spatial and/or temporal resolution is also necessary in certain kinds of studies (e.g., plume arrival times) and in comparisons of different models and methodologies. Integration or averaging of measurements over time or space often conceals useful detailed information on model performance. For these reasons, ARAC has emphasized a strict point-to-point protocol when pairing calculations to measurements to determine residuals, with no additional averaging beyond that inherent in the original supplied data.

No single statistical measure, or group of measures, provides a completely satisfying evaluation of model performance. Although previous analyses have included many of the common measures, historically a ratio factor analysis has been the principal method of performance summary due in part to its intuitively simple nature and its relatively straight-forward application to the combined results of experiments in which measurement values vary over orders of magnitude. However the use of ratios necessitates a decision concerning the applicability of zero-paired values (i.e., zero predicted or measured concentrations). In our protocol, only pairs of values in which either the measured or calculated value exceeds a specified threshold are used (the threshold is typically taken to be equal to, or slightly greater than, the measurement threshold). Each of these pairs is then classified into one of three groups:

1. non-zero measurements paired to non-zero calculated concentrations (expressed as ratios always greater than one),
2. non-zero measurements paired to zero value calculated concentrations, and
3. non-zero calculated concentrations paired to zero value measurements.

Each of the three groups are then analyzed separately.

Care is taken not to artificially inflate the statistical results. Group 1 ratios are consistently formed by first removing the background concentrations (if any) from the measurements and the percentages of ratios within a given factor are taken with respect to the number of all non-zero measurements (the combined number from groups 1 and 2). In this way the denominator used to form the factor percentages of a given data set remains constant from analysis to analysis. These procedures lead to relatively poorer statistical results than would result from alternative protocols typically used.

3. System Design and Current Implementation

The evaluation tools can be split into those emphasizing statistical measures of calculation-measurement residuals, and those emphasizing the visualization of model calculations, measurement data, and the relationship between the two. In practice, the statistical measures are used primarily for guidance in directing further analysis using the graphical visualization tools, and secondarily for performance summary.

Although not yet completed, significant portions of the system have been implemented. A key component is a *Data and Model Input Archive*. This archive stores the field experiment data needed to complete the evaluation studies. The archive is implemented as a distributed database and is designed to hold consistently-formatted meteorological and concentration measurements, information about the measurement sensors, and tracer source term characteristics. A user-interface is provided which allows data viewing and editing. Data currently processed and stored in this archive were taken from well over 100 separate tracer or other pollutant releases, and represent dispersion over a wide range of transport scales, topographical and meteorological complexity, atmospheric stability, and release characteristics. Initial selection of data for inclusion into the archive has been based on previous use in ARAC model evaluation studies or pertinence to specific interests (e.g., explosive release of particulate matter).

A *Statistics Calculations* subsystem provides the user with commonly-used statistical measures of the observation-calculation ratios and residuals (e.g., fractional or geometric bias, fractional scatter, correlation coefficient, normalized mean square error or geometric variance, etc.), as well as statistical descriptors of the measurement and calculation distributions (e.g., means, medians, variances, etc.). These measures may be applied, as appropriate, to any of the three groups of paired values described above.

Graphical visualization tools for exploratory data analysis have also been implemented (e.g., measured vs. calculated value scatter plots and cumulative frequency distributions), with additions planned for future implementation. Other, more tailored, displays are used to explore performance as a function of some quality or characteristic of the measurements, such as downwind distance and measurement rank (i.e., the measurements ordered from highest to lowest value). An example display of MATHEW/ADPIC results from the 10 1978-79 Copenhagen tracer experiments (Gryning, 1981) using 20-minute averaged air concentration measurements is shown in Fig. 1. Here the cumulative percentage of measurement/calculation ratios (always expressed as a number greater than 1) which fall within a factor of 2, 5, and 10 are plotted versus measurement rank (lowest 3 solid curves using the bottom axis labels) and versus distance from the source (lowest three dashed curves using the top axis labels). The highest dashed and solid curves indicate the number of group 1 measurements (i.e., those paired to non-zero calculations) expressed as percentages of the total number of non-zero measurements (100 minus this percentage is the percentage of measurements paired to zero-value calculations). The shaded histogram indicates the percentage of ratios within a factor of 2 for each decile of measurement rank. Also noted in the right margin (as M:C, M:0, and 0:C) are the relative sizes of each of the three pair groups mentioned above.

The dashed curves were derived by dividing the ratios into three sub-groups, those with measurements from the arcs within 3 kilometers of the source, those between 3 and 5 kilometers, and those greater than 5 kilometers. These results indicate a slight degradation of performance with distance from the source. The degradation due to measurement rank (solid curves and histogram) is more noticeable. This is a common general pattern across all data sets and usually indicates the difficulty in accurately predicting plume boundaries where measurements are near instrument threshold or background values and spatial concentration gradients are relatively large. Maximum MATHEW/ADPIC performance is usually, but not always, found for the highest 20-30% of the measurements.

To provide a more detailed understanding of model performance, a *Graphical Displays* subsystem provides graphical data comparisons geared for both objective and subjective evaluation. For example, although Fig. 1 reports the number of zero-paired values, it cannot be determined if these are due to model errors in transport direction, transport speed, and/or diffusion. To aid this kind of analysis, the system currently relies on displays such as that shown in Fig. 2. The bottom (left) window in Fig. 2 contains a visual comparison between the ADPIC modeled concentration contours and the tracer measurements for the first 20-minute period of the 26 September 1978 Copenhagen experiment. Measurement locations are depicted as black rings overlaid onto the calculated contours of tracer concentration. The colors (or gray scale values in this case) used for the contours represent the same concentration magnitude range as the corresponding color (or gray scale) used to fill those rings at non-zero measurement locations. (Therefore, the model matches the measurement to within the range represented by the color wherever a ring overlays a contour of the same color.) Among the additional displays the user may generate with this tool is a 2D concentration cross-section which is created by successively clicking on a series of measurement rings. This has been done for the outer arc of sampler rings (those rings connected with the black line), giving the measured and calculated concentration versus distance along that arc shown in the lower right corner of Fig. 2. (The curve with the higher peak represents the calculated plume.) Curves giving the measured and calculated values as a function of time at a given sampler location can also be generated.

In this example the contoured model calculations reproduce the measured plume magnitude and location reasonably well. A more detailed look at the cross-section produced from the outer sampler arc shows the model calculated the peak magnitude to within 10% and correctly located the peak's position along the arc. However the calculations do not reproduce the cross-wind spread of the plume at the arc's location, which leads to an underprediction across a significant portion of the

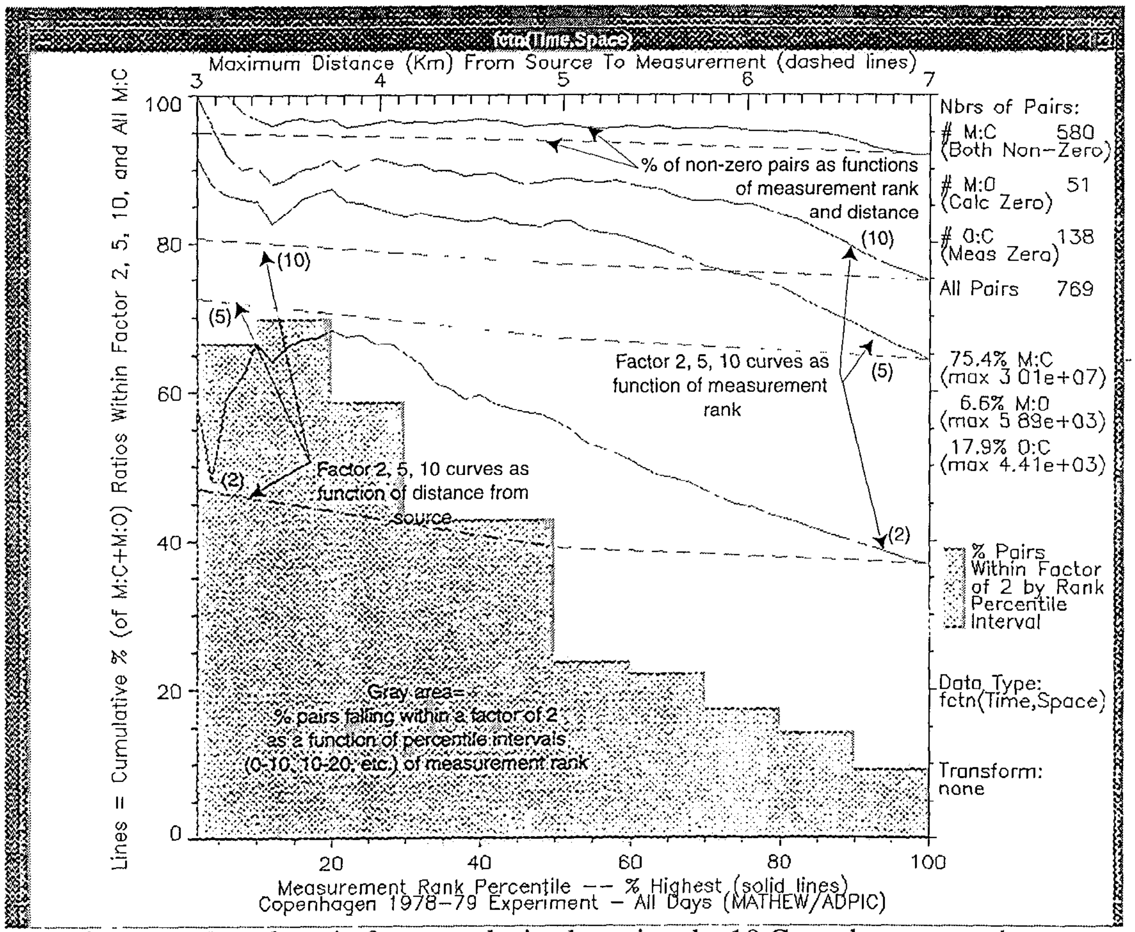
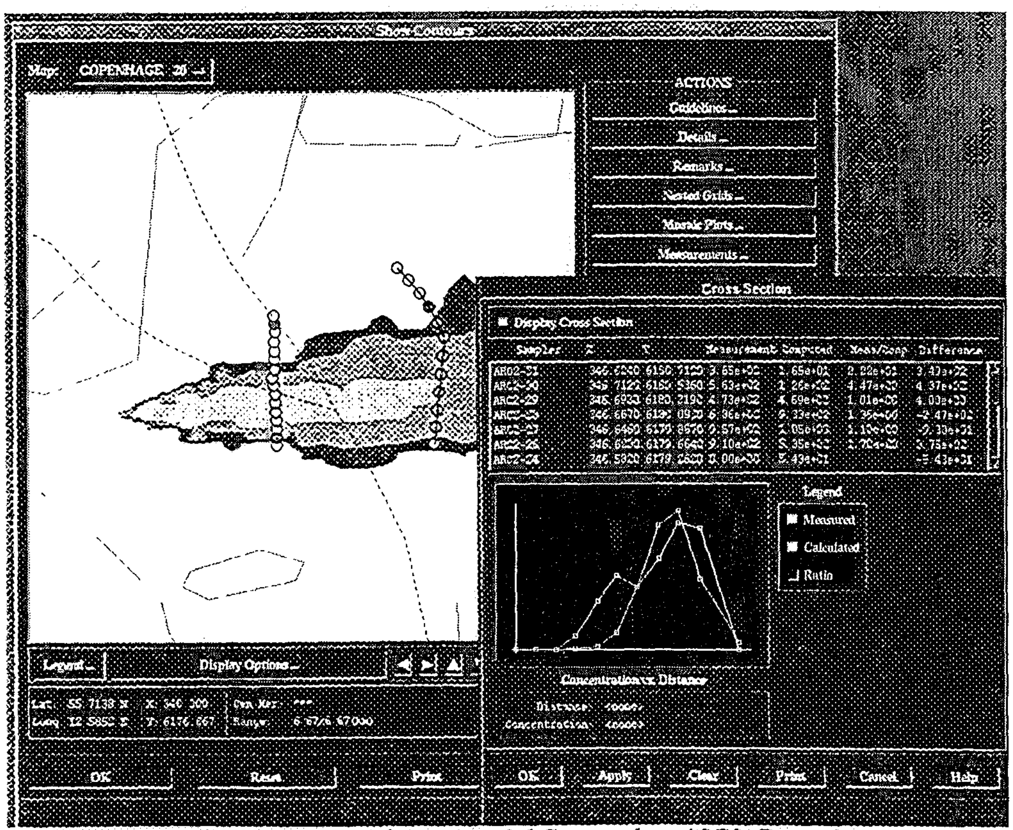


Figure 1. Sample ratio factor analysis plot using the 10 Copenhagen experiments



■ Display Cross Section

Sampler	X	Y	Measurement	Computed	Meas/Comp	Difference
ARC2-31	346.6280	6150.7120	3.65e+02	1.65e+01	2.22e+01	3.48e+02
ARC2-30	346.7122	6180.5360	5.63e+02	1.26e+02	4.47e+00	4.37e+02
ARC2-29	346.6933	6180.2190	4.73e+02	4.69e+02	1.01e+00	4.00e+00
ARC2-28	346.6670	6180.0930	6.86e+02	9.33e+02	1.36e+00	-2.47e+02
ARC2-27	346.6460	6179.8570	9.57e+02	1.05e+03	1.10e+00	-9.33e+01
ARC2-26	346.6250	6179.6640	9.10e+02	5.35e+02	1.70e+00	3.75e+02
ARC2-24	346.5920	6179.2520	0.00e+00	5.43e+01		-5.43e+01

Lat: 55.7138 N X: 340.300 Dom Max: ***
Long: 12.5852 E Y: 6176.667 Range: 6.67/6.67000

Distance: <none>
Concentration: <none>

Figure 2. Sample contour analysis plot using 26 September 1978 Copenhagen experiment.

plume. In this instance, an increase in the horizontal spread of the surface plume could lead to a better overall fit to the observations and might also eliminate the single zero-paired measurement illustrated at the northern (top) edge of the plume. Examination of similar products can be used to identify consistent trends. If such trends are found, the appropriate model algorithms, data, and assumptions may be re-examined to verify a proper implementation.

4. Results From Recent Model Performance Comparisons

As part of a major development program, ARAC's MATHEW and ADPIC operational codes will be replaced with new models currently under development. ADAPT (Sugiyama and Chan, 1998) is a much more general and flexible atmospheric data assimilation model than MATHEW. In generating mass-consistent windfields ADAPT offers a hierarchy of data assimilation methods, diverse interpolation and extrapolation techniques, as well as a variety of atmospheric parameterizations. LODI (Leone, et al., 1997) is a Lagrangian particle, random-walk dispersion model. Although similar in nature to ADPIC, LODI offers improved diffusion and deposition methodologies, greater flexibility in source description, and a more detailed spatially-varying description of the atmospheric boundary layer.

A key improvement in the ADAPT/LODI models is the use of a new gridding system based on a continuous terrain representation in which the surface is determined via a piecewise bilinear interpolation of gridded topographic data. The grids may have variable resolution in both the horizontal and the vertical and different resolutions may be used for the meteorological and concentration (particle sampling) grids. This allows improved resolution of topography, meteorological data, and particle sampling.

Operational evaluation of early development versions of ADAPT and LODI using the tools described above is currently underway. To date, initial benchmark calculations have been completed for 100 tracer release experiments using the following data sets:

1. Prairie Grass - 56 experiments of passive surface releases sampled on concentric arcs 50 to 800 meters downwind. Air samples were time-averaged for the entire plume passage.
2. Copenhagen - 10 experiments of passive elevated releases sampled on arcs from 2 to 6 kilometers downwind. Measurements were sequential 20-minute averaged air concentrations, taken over fairly flat, residential areas.
3. Mesoscale Atmospheric Tracer Experiments (MATS) - 25 experiments of elevated releases with momentum rise sampled on arcs from approximately 25 to 42 kilometers downwind. Measurements were sequential 7 to 22-minute averaged air concentrations, taken in rolling tree-covered terrain.
4. DOPPTEX - 8 experiments of surface and elevated passive releases sampled from approximately 1 to 61 kilometers downwind. Measurements were sequential 1-hour averaged air concentrations, taken in very complex coastal terrain.
5. Double Tracks - 1 experiment of explosively released particulate matter sampled on arcs from less than 1 to 15 kilometers downwind. Measurements were time-integrated air concentrations and accumulated deposition concentrations for the entire plume passage, taken in a broad valley.

This benchmark study provides a preliminary quantification of the statistical performance of ADAPT/LODI and allows both model development staff and prospective operational users to gain initial insights into their operational performance. Two points must be emphasized:

1. These benchmark results are from preliminary, pre-operational versions of ADAPT and LODI, and
2. Although ADAPT and LODI provide a much greater selection of parameter input and solution methods, input parameters and calculation algorithms selected for this initial set of runs mimic methods used in MATHEW and ADPIC. No analysis has yet been attempted to achieve better statistical results through the use of improved, or more applicable, methods now available or being developed in ADAPT/LODI. (Future studies will address these newer capabilities.)

Accordingly, differences in results are expected to be due to the inherent fundamental design differences between the two sets of models, such as the ADAPT/LODI continuous terrain representation vs. the discontinuous "stair-step" representation used in MATHEW/ADPIC.

Table 1 presents a summary of the ratio factor results for these initial benchmark runs. Statistical improvement is seen across all of the data sets for the ADAPT/LODI models. It should be noted that some of these results are based on very short measurement intervals which significantly reduce the percentages.

TABLE 1 Percentage of Ratios within Factor 2, 5, and 10 for Each Benchmark Data Set

	Prairie Grass % in Factor:			Copenhagen % in Factor:			MATS % in Factor:			DOPPTEX % in Factor:			Double Tracks* % in Factor:			Double Tracks** % in Factor:		
	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10
MATHEW/ADPIC	39	67	78	37	64	75	12	27	39	15	35	49	11	28	41	24	47	69
ADAPT/LODI	49	73	83	38	68	77	14	34	47	18	41	56	16	33	48	33	67	81
Difference in %	+10	+6	+5	+1	+4	+2	+2	+7	+8	+3	+6	+7	+5	+5	+7	+9	+20	+12

* Double Tracks air concentration measurements **Double Tracks deposition measurements

The ratio factor results provide a first impression of the performance differences. As an example of the use of the system for further investigation, Figs. 3a/3b show results from the MATS experiment conducted on 28 September 1983. The Savannah River valley runs from northwest to southeast through the center of the simulation region, with the plume traveling to the southwest. The contours show the calculated MATHEW/ADPIC and ADAPT/LODI contours for the last 15 minute period in which the sampler arc measured the plume passage. Comparison shows that the ADAPT/LODI plume travels faster than the one predicted by MATHEW/ADPIC. This can also be seen from the time history of the modeled and measured concentrations (see insets in Figs. 3a/3b) which are shown for the sampler marked with the black "X". The calculated values are represented by the curves having the higher peaks. The ADAPT/LODI calculations provide a significant improvement relative to MATHEW/ADPIC in plume departure times, peak concentration time (delay of 15 minutes rather than 30 minutes), and the ratio of calculated to measured peak values (unpaired in time) which has been reduced to 1.75 from 2.34. Other statistical measures emphasize the importance of the improvements in the overall advection speed. For example, a direct result of the improvement in plume departure time is a significant 30% reduction (795 to 562) in zero-paired calculated values for all 25 MATS experiments.

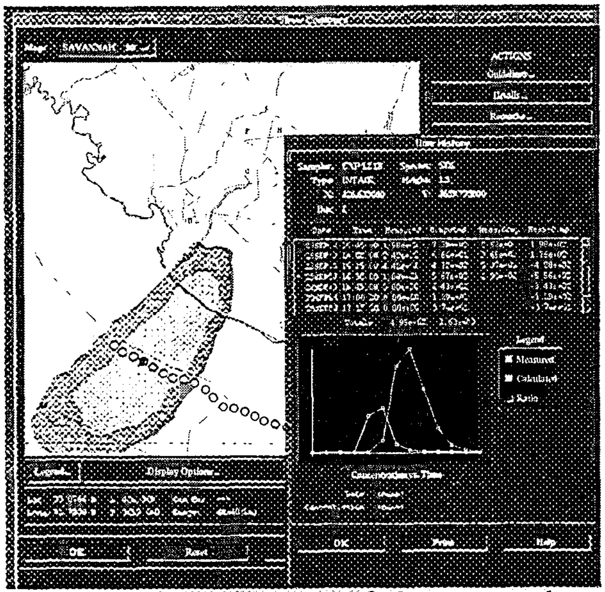


Figure 3a. MATHEW/ADPIC Contours and MATS Measurements.

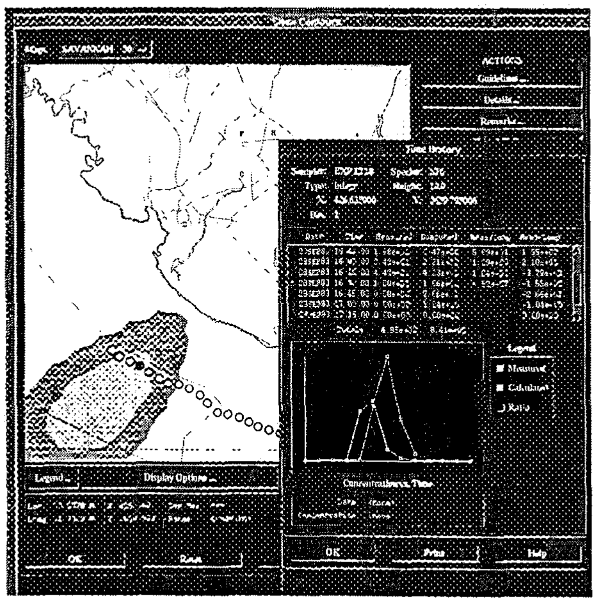


Figure 3b. ADAPT/LODI Contours and MATS Measurements.

Although a detailed analysis of the MATS experiments has not been completed, a likely cause of the improved ADAPT/LODI plume passage times and peak integrated concentrations is the use of a continuous terrain representation and improved near-ground resolution resulting in a more realistic windfield. Fig. 4 shows an east-west vertical cross-section through the lower portion of the ADAPT meteorological grid used for the MATS simulations. The grid uses variable vertical grading in a sigma-z coordinate, with the lowest level forming the terrain surface. The gray shaded area illustrates the discontinuous terrain representation used in the MATHEW/ADPIC models. The limited, fixed number of vertical levels in MATHEW necessitated a relatively coarse vertical cell resolution (35 meters in this case), whereas there is much higher resolution near the surface in the ADAPT grid.

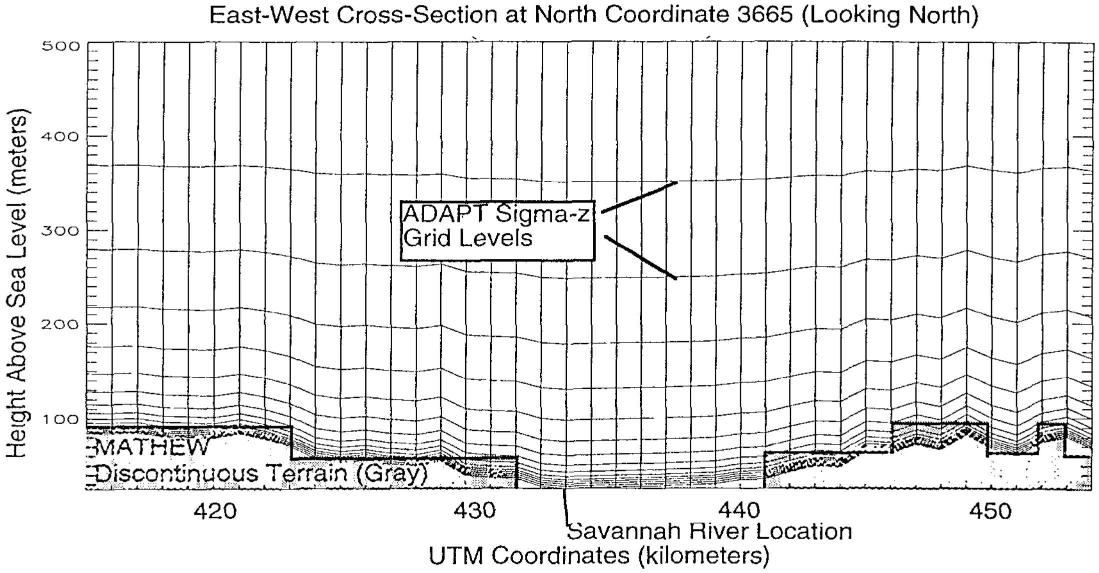


Figure 4. East-West Cross-Section of ADAPT Meteorological Grid with Overlaid Shaded MATHEW Terrain.

An examination of the near-surface MATHEW windfields (Fig. 5a) shows that advection speeds vary significantly within the Savannah River valley as the modeled flow moves over the discontinuous terrain. A general slowing of the winds causes particles to linger in the river valley. The speed reduction most likely stems from the mass-consistent adjustment process, which tends to reduce horizontal wind components when winds have a component perpendicular to vertical block topography faces. In contrast, Fig. 5b shows the corresponding ADAPT windfield. The smaller variation in vector length across the river valley indicates no apparent slowing of the ADAPT winds over the smoothly changing topography. (The MATHEW and ADAPT model graphics are generated using different graphics packages, causing the slight differences in vector and geography appearance in Figs. 5a and 5b.)

5 Summary

A prototype model evaluation system has been designed, and partially implemented. It will be employed to assess changes in performance of the operational production models used by the Atmospheric Release Advisory Capability (ARAC). The system may be used for both quality control procedures and for scientific investigation of new model features and algorithms. The system offers both objective statistical comparison methods and graphical visualization of the inter-relationships between model results and field measurement data. It has been applied recently to compare and benchmark pre-operational versions of ARAC's new core meteorological transport and diffusion models.

Acknowledgments

The authors would like to thank Hoyt Walker for developing the topographical data processing and model grid generation software used in this work.

This work was performed under the auspices of the U.S. Department of Energy by the Lawrence Livermore National Laboratory under contract No. W-7405-Eng-48.

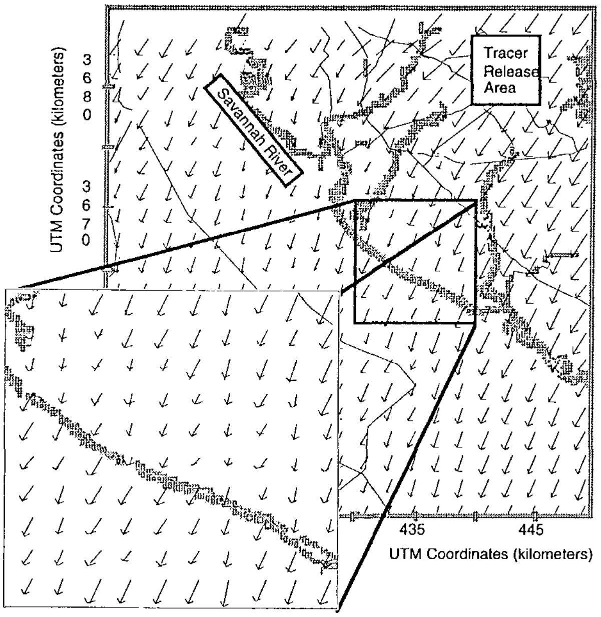


Figure 5a. MATHEW Windfield (28 September 1983 at 15:15 UTC) from MATS Run. Enlarged Area Shows Higher Windfield Resolution.

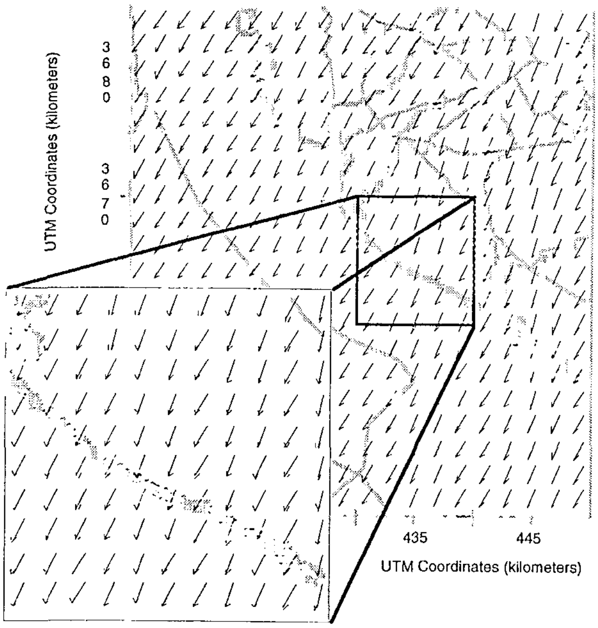


Figure 5b. ADAPT Windfield (28 September 1983 at 15:15 UTC) from MATS Run. Enlarged Area Shows Higher Windfield Resolution.

References

Ermak, D.L., J.S. Nasstrom, and A.G. Taylor (1995): Implementation of a Random Displacement Method (RDM) in the ADPIC Model Framework. Report UCRL-ID-121742, Lawrence Livermore National Laboratory, Livermore, CA.

Foster, K.T. and M.H. Dickerson (1990): An Updated Summary of MATHEW/ADPIC Model Evaluation Studies. Report UCRL-JC-104134, Lawrence Livermore National Laboratory, Livermore, CA.

Gryning, S.E. (1981): Elevated Source SF6 Tracer Dispersion Experiments in the Copenhagen Area. Report Risø-R-446, Risø National Laboratory, Denmark.

Lange, R. (1978): A Three-Dimensional Particle-In-Cell Model for the Dispersal of Atmospheric Pollutants and its Comparison to Regional Tracer Studies. *J. Appl. Meteor.*, **17**, 320-329.

Leone, J.M., Jr., J.S. Nasstrom and D.M. Maddix (1997): A First Look at the New ARAC Dispersion Model. *Proceedings of the ANS 6th Topical Meeting on Emergency Preparedness and Response*, San Francisco, CA, April 22-25, 1997.

Sherman, C.A. (1978): A Mass-Consistent Model for Wind Fields Over Complex Terrain. *J. Appl. Meteor.*, **17**. 312-319.

Sugiyama, G., R.L. Lee, and H. Walker (1994): Conjugate Gradient MATHEW. Report UCRL-ID-118629, Lawrence Livermore National Laboratory, Livermore, CA.

Sugiyama, G. and S.T. Chan (1998): A New Meteorological Data Assimilation Model for Real-Time Emergency Response. *Proceedings of the 10th Joint Conference on the Applications of Air Pollution with the Air and Waste Management Association*, Phoenix, AZ, January 11-16, 1998.

Sullivan, T.J., James S. Ellis, C.S. Foster, K.T. Foster, R.L. Baskett, J.S. Nasstrom, and W.W. Schalk, III. (1993) Atmospheric Release Advisory Capability: Real-time Modeling of Airborne Hazardous Materials. *Bull. Amer. Meteor. Soc.*, **74**, 2343-2361.